OKTOBER 2021

NÅR ALGORITMER SAGSBEHANDLER: KRAV TIL DATA I PROFILERINGSMODELLER BRUGT I DET OFFENTLIGE

Jo mere data en algoritmisk profileringsmodel bygger på, des bedre vil kvaliteten af dens afgørelser typisk være. Men når offentlige myndigheder bruger store mængder data om borgerne, stiller det høje krav til kvaliteten og nødvendigheden af oplysningerne. Dette notat belyser vigtigheden af høj datakvalitet og -beskyttelse, når offentlige myndigheder bruger profileringsmodeller.

INSTITUT FOR MENNESKERETTIGHEDER ANBEFALER SPECIFIKT:

- Justitsministeriet med inddragelse af Datatilsynet og Digitaliseringsstyrelsen udsteder en vejledning om offentlige myndigheders brug af profileringsmodeller med fokus på de rettigheds- og retssikkerhedsmæssige udfordringer ved modellerne.
- At der i vejledningen stilles krav om konsekvensanalyser for kunstig intelligens (AI-konsekvensanalyser) tidligt i beslutningsfasen og periodisk herefter under modellens udvikling og anvendelse til både beslutningsstøtte og fuldautomatisering.
- At konsekvensanalyserne bl.a. bør omfatte en vurdering af, om modellen overholder databeskyttelsesforordningens krav om dataminimering og sikrer brug af korrekte data,
- At der i vejledningen stiller krav om periodisk gentræning af modellen på nye data, eller at modellen udvikles således, at den som udgangspunkt giver mindre vægt til data, jo ældre de er.

Danmark er blandt de førende i verden, når det gælder digitalisering af borgernes kontakt med stat, region og kommune.[1] Digitaliseringen omfatter i stigende grad også brug af kunstig intelligens, hvor store mængder data selv "lærer" at analysere opgaver og finde løsninger. Ansigtsgenkendelse, maskinoversættelser og rutevejledninger er alle eksempler på brug af kunstig intelligens, som de fleste kender, men teknologien kan også bruges i offentlige myndigheders sagsbehandling, når de skal træffe afgørelser om for eksempel ydelser, indsatser eller sanktioner overfor borgere.

I 2019 udgav regeringen "National strategi for kunstig intelligens", hvori man understreger vigtigheden af fortsat udvikling på området.[2] Strategiens vision er, at "Danmark skal gå forrest med ansvarlig udvikling og anvendelse af kunstig intelligens", og visionen er allerede nu udmøntet i en række signaturprojekter[3] og forskningsprogrammer på både nationalt og EU-plan.

Trods det politiske fokus og de mange tiltag mangler der imidlertid et overblik over, hvordan offentlige myndigheders brug af kunstig intelligens griber ind i menneskerettighederne. Spørgsmålet er kun sporadisk blevet adresseret i Danmark, og det offentliges forskellige digitaliseringsstrategier kommer ikke nærmere ind på, hvilke følger det kan få for borgernes retssikkerhed, at dele af sagsbehandlingen udføres af algoritmiske profileringsmodeller.

Både i EU, Europarådet og FN[4] er der udtrykt bekymring for, at kunstig intelligens kan underminere beskyttelsen af den enkeltes rettigheder. Brugen af profileringsmodeller kan fx skabe risici for diskrimination på en række forskellige måder.

Institut for Menneskerettigheder anser offentlige myndigheders brug af kunstig intelligens for at have stor betydning for borgernes rettigheder, og instituttets arbejde tager netop udgangspunkt i en rettighedsbaseret tilgang. Dette har som forudsætning, at det skal være muligt at få indblik i myndighedernes brug af profileringsmodeller og til en vis grad også indblik i, hvordan de virker.

Instituttet vurderer overordnet, at der er behov for regler om, at offentlige myndigheder skal påvise, hvordan de sikrer samtlige rettigheder og retsgarantier for borgerne, når de bruger algoritmiske profileringsmodeller. Dette gøres efter instituttets vurdering bedst ved gennemførelsen af såkaldte 'artificial intelligence'-konsekvensanalyser (AI-konsekvensanalyser), som vi redegør for i et beslægtet notat.[5]

FOKUS FOR NOTATET

I dette notat belyser instituttet vigtigheden af, at de oplysninger, som den algoritmiske profileringsmodel trænes på og behandler i sagsbehandlingen, både er korrekte og nødvendige. Datakvaliteten og -minimering er afgørende for at sikre, at modellens afgørelser ikke fører til diskrimination eller usaglighed, og for at beskytte borgernes oplysninger. Det er således væsentligt, at krav til data sikres ikke blot i udviklingen af modellen, men også løbende gennem hele modellens levetid.

HVAD ER EN ALGORITMISK PROFILERINGSMODEL?

En algoritmisk profileringsmodel er en computerdrevet matematisk formel, som på baggrund af statistiske data eller variabler vurderer, om et objekt – eksempelvis en borger – har en bestemt egenskab, for eksempel om borgeren er berettiget til et tilskud.

Profileringsmodeller udvikles og trænes på baggrund af store mængder data om allerede afgjorte sager på området. Modellerne kan analysere og behandle data og finde sammenhænge og mønstre, der er langt mere komplekse, end hvad mennesker er i stand til, og de kan derfor være en stor ressource i sagsbehandlingen.

Profileringsmodeller kan både bruges til at træffe forvaltningsretlige afgørelser (fuldautomatisering) eller til at oplyse eller støtte myndighedens afgørelse (automatiseret beslutningsstøtte). Bruges modellen til beslutningsstøtte, kan dette enten ske for at afklare faktuelle omstændigheder som led i sagsoplysning, eller som støtte i den juridiske afvejning af en sag. Udfordringerne med transparens er til stede, uanset om modellen bruges fuldautomatiseret eller til automatiseret beslutningsstøtte.

Bruges modellen til beslutningsstøtte, kan dette enten ske for at afklare faktuelle omstændigheder som led i sagsoplysning eller som støtte i den juridiske afvejning af en sag. Udfordringerne med uigennemsigtighed er til stede, uanset om modellen bruges fuldautomatiseret eller til automatiseret beslutningsstøtte.

Profileringsmodeller er allerede under udvikling og i brug i det offentlige Danmark. Blandt andet har Styrelsen for Arbejdsmarked og Rekruttering udviklet og implementeret en model til forudsigelse af langtidsledighed blandt nyledige, og Gladsaxe Kommune udviklede i 2018 – men implementerede ikke – en model til forudsigelse af mistrivsel i børnefamilier. I begge tilfælde behandler modellen udvalgte data helholdvis om nyledige og om børnefamilier, og modellens resultater kan få betydning for den indsats, der igangsættes over for de identificerede ledige og børnefamilier.

KRAV TIL BRUGEN AF DATA I PROFILERINGSMODELLERNE

En profileringsmodel er bygget på og behandler oplysninger om borgere. Både mængden og kvaliteten af data har stor betydning for udviklingen af modellen og anvendelsen af den. I udviklingsfasen opbygges modellen på baggrund af historiske data, og i anvendelsesfasen behandler modellen løbende data om de borgere, som den skal vurdere.

Jo flere data – det vil sige oplysninger om borgerens forhold – en algoritmisk model trænes på, desto højere bliver modellens kvalitet typisk, fordi modellen

bliver bedre i stand til at foretage komplekse og nuancerede vurderinger. Sammenhængen mellem datamængde og kvalitet kan skabe en stærk tilskyndelse til at lade algoritmen behandle så store mængder data som muligt.

Myndighederne er imidlertid ikke frit stillede på dette område: brugen af oplysninger om borgere er reguleret i både forvaltningsretten og i databeskyttelsesforordningen, og disse regler sætter rammerne for mængden og kvaliteten af data, der må bruges.

Af forvaltningsretten fremgår det, at myndighederne skal oplyse en sag i nødvendigt omfang og sikre, at den data, der anvendes, er korrekt. Princippet kaldes officialprincippet og har afgørende betydning for profileringsmodellens lovlighed, fordi unødvendige eller forkerte data kan få betydning for modellens evne til at træffe korrekte afgørelser.

Databeskyttelsesforordningen har et princip om dataminimering. Det har til formål at sikre, at brugen af personoplysninger ikke bliver uforholdsmæssigt indgribende. Dataminimeringsprincippet indebærer, at uforholdsmæssigt store mængder data ikke må behandles, og at en hvilken som helst oplysning, der ville føre til et uproportionalt indgreb, heller ikke må behandles.

Fælles for officialprincippet og dataminimeringsprincippet er, at de rummer vurderinger af, om de indsamlede oplysninger om borgeren er nødvendige i forhold til det formål, som myndigheden forfølger.

Når offentlige myndigheder beslutter at introducere en profileringsmodel i sagsbehandlingen anbefaler vi, at der tidligst muligt gennemføres en såkaldt AI-konsekvensanalyse, der specifikt forholder sig til kravene om blandt andet dataminimering. Disse analyser bør desuden gennemføres periodisk i hele modellens levetid. Vi beskriver konsekvensanalyserne nærmere i et beslægtet notat.[6]

RISIKO FOR DISKRIMINATION OG USAGLIGHED HÆNGER SAMMEN MED DATAKVALITET

Selve udvælgelsen af data til træning af modellen – for eksempel tidligere afgørelser fra myndigheden – er væsentlig, fordi datasættet kommer til at udgøre modellens ”grundsandhed”.

Forholdet mellem kvaliteten af data og modellens vurderinger opsummeres ofte i sloganet: ”Garbage in, garbage out”. Det skal forstås på den måde, at hvis tidligere afgørelser er behæftet med fejl, vil der være risiko for, at modellen viderefører de fejl. Det samme gælder, hvis man træner modellen på data, der ikke er korrekte eller retvisende for den del af befolkningen, som berøres af modellen. Her er risikoen, at modellen trænes på usaglige eller diskriminerende – og dermed ulovlige – sammenhænge.

I udviklingsstadiet skal der derfor være fokus på, hvordan data udvælges, så det både er korrekt og retvisende for samtlige befolkningsgrupper. Hvis der korrigeres eller forsimples i datasættet, skal datasættet stadig have samme kvalitet.

Datakvalitet kan ligeledes forringes, hvis data gradvist forældes. Det skyldes, at omstændighederne forandrer sig, så de sammenhænge, som det oprindelige datasæt indeholder, ikke længere passer med virkeligheden. Opdatering af datasættet kan forbedre datakvaliteten, især hvis myndigheden opdaterer data med fokus på, at forkerte data rettes til eller udgår fra datasættet. Man kan forsøge at løse udfordringen med forældede data ved periodisk at gentræne modellen på et datasæt, der kun indeholder nyere oplysninger, eller ved at få algoritmen til at vægte ældre data mindre.

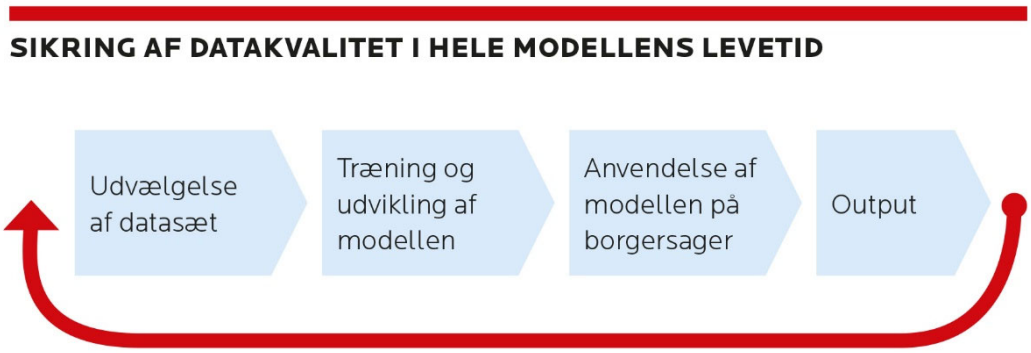
SIKRING AF DATAKVALITET I HELE MODELLENS LEVETID

Hvis myndigheden ikke opdager fejl i datasættet, kan det føre til såkaldte negative feedbacksløjfer, der kan forstærke risikoen for usaglighed og diskrimination. I en negativ feedbacksløjfe trænes modellen på data, der fører til enten usaglighed eller diskrimination, og som modellen så viderefører i sine egne vurderinger. Vurderingerne indgår i en myndigheds praksis, som derfor også kommer til at indebære usaglighed eller diskrimination i afgørelsen. Når nyere praksis herefter inkluderes i datasættet for en opdateret model, får modellen tendens til at foretage en endnu stærkere forskelsbehandling (se figur på næste side).

En negativ feedbacksløjfe vil opstå, når det er vanskeligt at måle eller opdage forkerte resultater. Anvendes en profileringsmodel for eksempel til tidlig opsporing af børns mistrivsel, skal modellen vurderes både i forhold til, hvor mange sager modellen opdagede, der var korrekte, men også om der var tilfælde, som burde være opdaget, men som modellen vurderede, at der ikke var grund til at fokusere på. Det sidste er selvsagt vanskeligt.

En måde at forhindre eller begrænse virkningen af negative feedbacksløjfer er at introducere et element af tilfældighed i afgørelsesprocessen.[7] Den negative feedbacksløjfe opstår, fordi modellen påvirker beslutningsprocessen og opdateres med data om resultaterne af disse beslutninger.

Ved at introducere tilfældighed i processen sikrer man, at modellen også opdateres med data, som er skabt uafhængigt af modellens vurderinger. Derved får den løbende træning af modellen mulighed for at korrigere de effekter, som en negativ feedbacksløjfe skaber. Et tænkt eksempel kunne være en model, som anbefaler selvangivelser til SKAT, som skal udtages til manuel kontrol. SKAT kunne i den situation forsøge at forebygge en negativ feedbacksløjfe ved også at udtage en vis kvote af tilfældigt udvalgte selvangivelser til manuel kontrol.



Udfordringen for metoden med at introducere tilfældighed i beslutningsprocessen er, at dette i nogle situationer vil være retssikkerhedsmæssigt problematisk.

Et element af tilfældighed er relativt uproblematisk i eksemplet med udtagning af selvangivelser til manuel kontrol. Som et eksempel på det modsatte kan nævnes, at mange stater i USA benytter modeller til at vurdere risici i forbindelse med beslutninger om varetægtsfængsling og prøveløsladelse.[8] Introduktionen af tilfældighed i en sådan beslutning, således at visse sigtede eller dømte blev løsladt eller fængslet fordi de var tilfældigt udtrukket, ville være indlysende uacceptabelt. Det samme kunne være tilfældet i særligt indgribende forvaltningsafgørelser truffet af myndighederne.

EU-Kommissionen har præsenteret et udkast til en forordning om kunstig intelligens. I den foreslår kommissionen, at AI-systemer skal udvikles på en måde, så virkningerne af eventuelle negative feedbacksløjfer adresseres for at mindske deres virkning i systemet.[9] Det bør efter vores vurdering ske som led i periodiske AI-konsekvensanalyser.[10]

ANBEFALINGER: HUSK RETTIGHEDERNE

Danmark er verdensmester i digital forvaltning. Offentlige myndigheder er ambitiøse, når det gælder udviklingen af digitale løsninger – men de har ikke taget stilling til, hvordan disse løsninger skal afspejle den fortsatte sikring af menneskerettigheder og retsprincipper.

En rettighedsbaseret tilgang til offentlige myndigheders brug af profileringsmodeller er særlig vigtig, da det gælder afgørelser om borgere. Der kan være tale om afgørelser, som har stor indvirkning på borgernes retsstilling som for eksempel afgørelser om udsatte borgeres behov for støtte. Offentlige myndigheder er forpligtet til at handle sagligt og til at sikre beskyttelsen af borgernes menneskerettigheder, også når sagerne automatiseres.

Det er afgørende, at offentlige myndigheder ikke uforvarende forstærker eksisterende uligheder ved at implementere teknologi, der ikke kan tage højde for rettighedernes fulde anvendelsesområde.

Institut for Menneskerettigheder efterlyser et større fokus på de mange rettighedsmæssige udfordringer, den digitale forvaltning stiller os overfor. Det indebærer klarere rammer for myndighedernes anvendelse af profileringsmodeller, både når det gælder datakvalitet og test, men også når det gælder risici og begrænsningerne i, hvad modellerne er i stand til.

LÆS MERE I VORES RAPPORT

I rapporten 'Når algoritmer sagsbehandler – Rettigheder og retssikkerhed i offentlige myndigheders brug af profileringsmodeller' udgivet i oktober 2021 kortlægger Institut for Menneskerettigheder, hvordan borgernes rettigheder og retssikkerhed påvirkes, når offentlige myndigheder benytter sig af kunstig intelligens. Rapporten er en af de første, der tager udgangspunkt i danske myndigheders praksis og dansk lovgivning.

Dette notat gengiver dele af rapportens konklusioner.

Læs hele rapporten på menneskeret.dk/algoritmer.

SLUTNOTER

[1] Digitalstyrelsen (2020), Ny FN-måling: Danmark er fortsat verdensmestre i offentlig digitalisering, tilgængelig på: https://digst.dk/nyheder/nyhedsarkiv/2020/juli/ny-fn-maaling-danmark-er-fortsat-verdensmestre-i-offentlig-digitalisering/

[2] Finansministeriet og Erhvervsministeriet (2019): National strategi for kunstig intelligens tilgængelig på: https://www.regeringen.dk/media/6537/ai-strategi_web.pdf

[3] Digitalstyrelsen (2019), Kommuner og regioner skal afprøve kunstig intelligens for at løfte kvaliteten i den offentlige service, tilgængelig på: https://digst.dk/nyheder/nyhedsarkiv/2019/oktober/kommuner-og-regioner-skal-afproeve-kunstig-intelligens-for-at-loefte-kvaliteten-i-den-offentlige-service/

[4] Kommissionens hvidbog (2020), " On Artificial Intelligence - A European approach to excellence and trust" Tilgængelig på: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf , Europarådets resolution, CM/Rec(2020)1, tilgængelig på: https://rm.coe.int/09000016809e1154 , FNs Højkommissariat for Menneskerettigheder (2019), "FN's Specialrapportør for ekstrem fattigdom", rapport A/74/48037.

[5] Institut for Menneskerettigheder, (2021) "Når algoritmer sagsbehandler: Klare rammer for det offentliges brug af profileringsmodeller", https://menneskeret.dk/udgivelser/klare-rammer-offentliges-brug-profileringsmodeller

[6] Institut for Menneskerettigheder, (2021) "Når algoritmer sagsbehandler: Klare rammer for det offentliges brug af profileringsmodeller", https://menneskeret.dk/udgivelser/klare-rammer-offentliges-brug-profileringsmodeller

[7] Kroll, Huey, Barocas, Felten, Reidenberg, et al. (2017). "Accountable Algorithms." University of Pennsylvania Law Review 165(3): 633

[8] Berk, Heidari, Jabbari, Kearns and Roth (2018). "Fairness in Criminal Justice Risk Assessments: The State of the Art." Sociological Methods & Research

[9] Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final artikel 15

[10] Institut for Menneskerettigheder, (2021) "Når algoritmer sagsbehandler: Klare rammer for det offentliges brug af profileringsmodeller", https://menneskeret.dk/udgivelser/klare-rammer-offentliges-brug-profileringsmodeller